Программа учебной дисциплины «Базовые методы анализа данных и работа со статистическими пакетами»

Утверждена

Академическим советом ООП

Протокол № 1 от «30» августа 2019 г.

Автор	Хавенсон Т.Е., Чиркина Т.А.
Число кредитов	7
Контактная работа (час.)	96
Самостоятельная работа (час.)	170
Курс	1 курс магистратуры
Формат изучения дисциплины	blended learning

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Базовые методы анализа данных и работа со статистическими пакетами» являются:

- сформировать у студентов теоретические представления об основных современных методах анализа данных в социальных науках;
- выработать навыки практического применения методов, как к самостоятельно собираемым данным, так и к базам данных;
- выработать у студентов представления о том, какие теоретические модели заложены в различных методах анализа данных;
- сформировать умение сопоставлять эти модели с задачами конкретного исследования и правильно выбирать метод в соответствии с его целями, задачами, гипотезами и имею- щимися данными.

В результате освоения дисциплины студент должен:

Знать

- основные теоретические представления о современных методах анализа данных в психологии и исследованиях в области образования;
- правила применения современных методов анализа данных в социальных науках;
- основные методы одномерной статистики и алгоритмы их реализации в пакете SPSS/R;
- основные методы многомерной статистики и алгоритмы их реализации в пакете SPSS/R;
- основные принципы визуализации получаемой в ходе анализа данных информации;

- основные источники информации, необходимые для реализации исследований (базы данных, тематические веб-сайты, главные научные журналы по данной тематике);
- принципы написания аналитических текстов.

Уметь

- реализовывать основные одномерные и многомерные методы анализа данных;
- анализировать специфику использования методов математики и статистики для изучения психологических и образовательных явлений;
- интерпретировать результаты анализа данных, полученных в ходе исследований в психологии и в области образования;
- ставить задачи для анализа данных в различных ситуациях в зависимости от типа данных и от исследовательских задач;
- учитывать ограничения различных методов анализа данных, оценивать качество полученной эмпирической информации;
- представлять результаты исследований для разных аудиторий слушателей;
- критически анализировать информационные источники, научные тексты, результаты других исследований;
- участвовать в проектных формах работы;
- реализовывать самостоятельные аналитические проекты;
- подкреплять свою точку зрения знаниями, полученными на основе статистического анализа существующих или собранных специально для этих целей данных;
- работать в пакете для статистической обработки данных SPSS/R и специализированной программе для обработки данных международных сравнительных исследований в области образования IDB Analyzer (IEA).

Владеть

- Навыками работы в программной оболочке для обработки и анализа информации на компьютере – SPSS/R;
- Навыками расчета простых статистических показателей вручную;
- Навыками практического применения методов к самостоятельно собираемым данным;
- Навыками практического применения методов к существующим базам данных;
- Навыками участия в проектных формах работы;
- Навыками реализации самостоятельных аналитических проектов;
- Навыками написания аналитических текстов с применением результатов анализа данных.

Изучение дисциплины «Базовые методы анализа данных и работа со статистическими пакетами» базируется на следующих дисциплинах:

- Школьный уровень владения математикой;
- Параллельное освоение курса «Теория и методология современной психологии: принципы измерений в психологии и образовании».

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать основные понятия и теории математики;
- знать простейшие методы решения математических задач;

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- Качественные и количественные методы исследований в психологии;
- Теория и практика разработки контрольно-измерительных материалов;
- Психометрические теории и анализ тестовых заданий;
- Национальные и международные программы оценки образовательных достижений;
- Курсовая работа и выпускная квалификационная работа.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Общие представления об анализе данных

Вводная тема, целью которой является общее погружение студентов в проблематику исследований в социальных науках в общем, и в психологии и образовании в частности. Обсуждаются основные типы исследований, цели анализа данных. Соответствие определенных методов анализа данных целям и задачам исследования. Разработка стратегии анализа данных. Процесс анализа данных в исследовании.

Тема 2. Основы теории вероятностей и математической статистики

Классическое и статистическое определение понятия вероятности. Правила сложения и умножения вероятностей. Формула полной вероятности. Условная вероятность. Зависимость/независимость событий.

Тема 3. Основные виды распределений вероятностей

Случайные величины - дискретные и непрерывные. Общие представления о распределении вероятностей. Нормальное распределение.

Произвольное и стандартное нормальное распределение. Стандартизация. Работа с таблицами нормального распределения. Квантили распределения.

Работа с таблицами других статистических распределений (распределение Стьюдента, хи- квадрат распределение, распределение Фишера), нахождение критических значений для проверки статистических гипотез.

Тема 4. Основы работы с пакетом SPSS / R

Начиная с 2018-2019 учебного года, в рамках практических занятий слушателям предлагается самостоятельно выбрать трек, по которому они будут осваивать реализацию методов: анализ в пакете SPSS (упрощённый) или анализ в пакете R (усложнённый). На практических занятиях будут использоваться раздаточные материалы в обоих пакетах. При этом задания, базы данных, отработка материала теоретических занятий идентична, как и получаемые результаты анализа.

Выбор потока происходит добровольно, без пререквизитов для усложнённого трека. Однако общей рекомендацией будет выбирать трек с пакетом R только для тех, у кого освоение теоретического материала не будет занимать много времени, то есть в случае, когда по нему есть предварительные знания.

Начало работы в SPSS. Правила создания макета анкеты (опросного документа). Ввод данных. Работа с переменными – кодирование, автоматическое и ручное, вычисление новых переменных, свойства переменных. Функции Recode, Count, Compute и д.р. Работа с данными – сортировка, отбор случаев, извлечение случайной выборки, агрегирование, чистка данных. Функции Sort cases, Select cases, Aggregate, Split file и д.р. Работа с файлами – слияние нескольких файлов, экспорт и импорт данных. Функции Merge file и д.р.

Начало работы с R. Загрузка и подключение пакетов, работа с Data Frame, открытие и сохранение данных. Элементы синтаксиса. Перекодировка, сортировка данных, создание новых переменных, отбор подвыборок. Изучается самостоятельно с помощью курса «Анализ данных в R» на платформе Stepik или с помощью курса «Введение в R» на платформе DataCamp теми студентами, которые выбрали трек с R.

Тема 5. Методы описательной статистики. Визуализация данных

Одномерные частотные таблицы, абсолютные и относительные частоты (процент, доля), накопленная частота. Основные типы шкал и соответствующие им меры средней тенденции и меры разброса. Линейное и нелинейное преобразование шкал. Стены, станайны. Точечное и интервальное оценивание.

Принципы графического представления данных. Наиболее популярные виды графиков: гистограмма, диаграмма рассеивания, диаграмма «ящик с усами» и др.

Правила оформления результатов описательной статистики.
SPSS: Descriptive Statistics: Descriptives, Explore, Q-Q plots, Frequencies. Graphs R: Описательная статистика, базовые графики, функция ggplot.

Тема 6. Анализ связи между двумя признаками

Таблицы сопряженности. Возможное содержание ячеек таблицы. Условные и

безусловные частоты. Повтор правила умножение вероятностей.

Коэффициенты парной связи для различных типов шкал. Критерий Хи-квадрат и основанные на нем коэффициенты. Коэффициенты корреляции. Проверка значимости корреляционной связи. Работа с таблицами множественных ответов.

Правила оформления результатов.
SPSS: Correlate, Crosstabs, Custom tables, Multiple Response

R: chisq.test, binom.test, fisher.test, cor.test

Тема 7. Общие принципы проверки статистических гипотез (параметрические и непараметрические критерии)

Общие правила проверки статистических гипотез. Нулевая и альтернативные гипотезы, p-value.

Алгоритмы проверки наиболее важных гипотез. Гипотезы о равенстве средних: тесты для одной выборки (z-test, t-test) и двух выборок (зависимые и независимые). Гипотеза о равенстве долей. Проверка значимости коэффициента корреляции. Доверительный интервал и уровень значимости. Ошибки 1 и 2 рода. Проверка нормальности.

Цели применения непараметрических методов. Работа с малыми выборками. Непараметрические критерии: критерии t-тест Манна-Уитни, W-тест Уилкоксона.
SPSS: Compare means (One-Sample T Test, Independent-Samples T Test)

R: t-tests, wilcox.test

Тема 8. Дисперсионный анализ

Формальная модель, заложенная в методе. Одномерный и многомерный дисперсионный анализ. Множественные сравнения. Интерпретация результатов.

Изучается самостоятельно с помощью курсов Сравнение средних и создание групп/Inferential Statistics.

SPSS: Compare means (Means, one-way anova)
R: функция aov и ее конфигурации, model.tables, ggplot

Тема 9. Линейный регрессионный анализ

Цели применения регрессионных моделей. Парный и множественный линейный регрессионный анализ. Выбор зависимых и независимых признаков и оценка качества построенной модели. Интерпретация коэффициентов регрессии. Регрессия с фиктивными переменными. Ограничения линейной регрессии.

Правила оформления результатов. SPSS: Regression (Linear)

R: функция lm, predict и их конфигурации, ggplot

Изучается самостоятельно с помощью курсов Основы статистики. Часть 3 и Inferential Statistics

Тема 10. Логистическая регрессия

Логистическая регрессия. Оценка качества полученной модели. Интерпретация полученных результатов.

SPSS: Regression (Binary Logistic)
R: функция prediction, performance и их конфигурации (package ROCR), ggplot

Тема 11. Факторныйанализ

Метод главных компонент. Цели применения метода. Алгоритм проведения анализа, требования к исходным данным, факторные нагрузки, вращение осей. Интерпретация результатов.
SPSS: Dimension reduction (Factor)

R: функции as.formula, factanal

Тема 12. Методы классификации признаков. Кластерный анализ и деревья классификации.

Основные цели, решаемые кластерным анализом, сфера применения. Иерархический кластерный анализ. Способы вычисления расстояний между объектами. Способы кластеризации.

Неиерархический кластерный анализ, метод k-средних. Совместное применение иерархических и неиерархических методов кластеризации.
SPSS: Classify (hierarchical, k-means cluster)

R: Функции hclust, kmeans

Тема 13. Написание отчета по проведенному анализу данных

Работа с окном выдачи результатов – SPSS Viewer. Редактирование таблиц, графиков. Экспорт объектов в приложения MS Office.

Общие принципы представления результатов применения статистических методов анализа данных. Отбор релевантной информации. Визуализация информации.

III. ОЦЕНИВАНИЕ

Оценка за текущий контроль в 1 и 2 модулях учитывает результаты студента по следующим компонентам:

- Домашние задания – 5 баллов
- Активность на занятиях (мини-тесты) - 1 балл
- Проектная работа – 2 балл
- Контрольная работа – 2 балла

Оценка за текущий контроль в 3 модуле учитывает результаты студента по следующим компонентам:

- Домашние задания – 5 баллов
- Контрольная работа – 2 балла
- Проектная работа (доклады, проекты) – 3 балла.

Итоговая оценка курс:

1) Текущая оценка за 1-3 модуль – 6 баллов
2) Экзаменационная работа – 4 балла

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Примерные вопросы/ задания для домашнего задания

Пример 1. Найдите достаточно простую русскоязычную статью (отчет по описательному исследованию, например, http://cim.hse.ru/unimonitoring), в которой изложены результаты исследования, включая описание процесса сбора и анализа данных и интерпретацию результатов.

Напишите комментарий к статье (1-2 страницы, до 4000 знаков). По примерному плану:

Название рецензируемого материала. Полное библиографическое описание.

1. Цель исследования, основной исследовательский вопрос или вопросы.
2. Методы сбора данных. Выборка.
3. Какие методы анализа данных применялись для получения результатов.
4. Основные полученные выводы и результаты.
5. Какой информации Вам не хватает в публикации.
6. Какие вопросы Вы задали бы авторам исследования.
7. Как бы Вы построили исследование с тем же исследовательским вопросом.

В выполненной работе необходимо указать либо полное библиографическое описание статьи, либо ссылку на сайт с материалами. Если это статья не из интернет-источника, желательно приложить текст статьи.

Пример 2. Описательная статистика. Работа выполняется на базе TIMSS 2015 (по студентам) по вашей стране. Необходимо выбрать несколько переменных разного уровня измерения и провести анализ методами описательной статистики.

Построить частотное распределение, найти меры центральной тенденции и разброса, описать полученные ре- зультаты для нескольких переменных (как минимум по 1 переменной для каждой шкалы). Если для переменной воз- можно найти несколько показателей, запишите все

Оценочные средства для промежуточной аттестации

1. Модели независимости, заложенных в коэффициентах, основанных на критерии хи-квадрат, коэффициентах ранговой корреляции, коэффициенте корреляции Пирсона
2. Т-Тесты :для независимых выборок, для парных выборок, одновыборочный (проверяемая стат.гипотеза. алгоритм проверки, содержательные выводы).
3. Дисперсионный анализ (проверяемая стат.гипотеза. алгоритм проверки, содержательные выводы).
4. Линейная регрессионная модель: оценка качества модели и значения коэффициентов регрессионного уравнения, ограничения
5. Линейная регрессионная модель: стандартизованные коэффициенты регрессионного уравнения
6. Бинарная и множественная логистическая регрессия: оценка качества модели, значение коэффициентов
7. Алгоритм иерархического агломеративного кластерного анализа
8. Кластерный анализ: алгоритм К-средних
9. Разведывательный факторный анализ. Метод главных компонент

V. РЕСУРСЫ

5.1 Основная литература

1. Tabachnick, B. G., & Fidell, L. S. (2012). Using multivariate statistics. Pearson.

Онлайн учебник с описанием реализации в разных пакетах, с аннотированными синтаксисами и т.п.

2. http://www.ats.ucla.edu/stat/

Каналы с обучающими видео по SPSS для начинающих:

3. http://www.youtube.com/playlist?list=PL5DE86FDC53716BEE
4. http://www.youtube.com/user/statisticsfun

Материалы для онлайн части курса:

1. Сравнение средних и создание групп, Новосибирский государственный университет. Платформа Coursera,
2. Основы статистики. Часть 3. Платформа Stepik, СПБГУ
3. Анализ данных в R. Часть 1, Платформа Stepik, СПБГУ
4. Quantitative methods, платформа Coursera, University of Amsterdam
5. Inferential Statistics, платформа Coursera, University of Amsterdam
6. Введение в R, платформа DataCamp

5.2 Программное обеспечение

п/п	Наименование	Условия доступа
1.	Microsoft Windows 10	*Из внутренней сети университета (договор)*
	SPSS	*Из внутренней сети университета*

.		*(договор)*
.	R	*Свободное лицензионное соглашение*

5.3 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.